Remarks

Claims 1-4, 6-11, 13, 16 and 41-44 have been amended. Upon entry of the foregoing amendments, claims 1-16, and 41-44 are pending. No new matter is added by these amendments. Support for the amendments may be found in the original claims and throughout the specification, *e.g.*, at page 17, lines 17–20; page 18, lines 11-15; page 21, line 1 through page 26, line 26; page 46, line 1 through page 48, line 10; and Example 2.

Applicant thanks the Examiner for returning a copy of the initialed Form 1449, which was submitted with Applicant's Reply to Office Action filed on February 27, 2003.

I. Rejections under 35 U.S.C. § 112, First Paragraph (Enablement)

Claims 1-16 and 41-44 stand rejected under 35 U.S.C. § 112, first paragraph, as containing subject matter that was not described in the specification in such a way as to enable one skilled in the art to which it pertains, or with which it is most nearly connected, to make and/or use the invention. Office Action at page 2. The Office alleges that "[w]hile the specification provides some guidance for a method of determining a probability value for the above listing using the particular equations or values disclosed, the specification does not provide guidance for a method of determining probability by any other means." Office Action at page 3. The office further alleges that "[g]iven the lack of descriptive working examples in the specification, and the unpredictability of generating probability values, the specification as filed is not enabling for any method of determining the listed probability values as claimed. The instant application is only enabled for the above-mentioned computational means of the four probabilities." Office Action at page 3. Applicant respectfully disagrees.

Applicant thanks the Examiner for acknowledging that the specification is enabling for the following equations: initial oligonucleotide probability (p. 21, equation I), transition probability (p. 22, equation II), nucleic acid sequence probability (p. 23, equation III), and probability of each state for the nucleic acid sequence (p. 24, equation IV). Office Action at page 3. Applicant respectfully disagrees, however, with the

Office's allegation that the specification does not enable a person skilled in the art to practice the invention commensurate in scope with the claims.

Disclosure of a single species provides sufficient enabling support if one of skill in the art can, using the state of the art and Applicant's written disclosures, practice the invention in its full scope without undue experimentation.[1] *See In re Wands*, 858 F.2d 731, 737, 8 U.S.P.Q.2d 1400, 1404 (Fed. Cir. 1988); *John Hopkins Univ. v. Cellpro, Inc.*, 152 F.3d 1342, 1361, 47 U.S.P.Q.2d 1705, 1719 (Fed. Cir. 1998) (Applicant's specification provided sufficient enabling support for the Applicant's claim to immunoassay methods using a generic class of antibodies even though Applicant made a public deposit of only a single hybridoma cell line that secreted a specific antibody); *Spectra-Physics, Inc. v. Coherent, Inc.*, 827 F.2d 1524, 1533, 3 U.S.P.Q.2d 1737, 1743 (Fed. Cir. 1987), *cert. denied,* 484 U.S. 954 (1987). Section 2164.03 of the M.P.E.P. states that "[a] single embodiment may provide broad enablement in cases involving predictable factors,[2] such as mechanical or electrical elements." *Citing In re Vickers,* 141 F.2d 522, 526-27, 61 U.S.P.Q. 122, 127 (C.C.P.A. 1944); *In re Cook*, 439 F.2d 730, 734, 169 U.S.P.Q. 298, 301 (C.C.P.A. 1971). Furthermore, it is well established law that patent applicants are not required to disclose every species enabled by their claims. *See In re Vaeck*, 947 F.2d 488, 496, 20 U.S.P.Q.2d 1438, 1445 (Fed. Cir. 1991).

Applicant need only show that one skilled in the art would be able to make and use the claimed invention using the application as a guide. *In re Brandstadter*, 484 F.2d 1395, 1406-07, 179 U.S.P.Q. 286, 294 (C.C.P.A. 1973). In order to be enabling, the

---

[1] Applicant notes that the performance of routine and well-known steps cannot create undue experimentation even if it is laborious. *See In re Wands*, 858 F.2d at 737, 8 U.S.P.Q.2d at 1404; *In re. Angstadt,* 537 F.2d 498, 504, 190 U.S.P.Q. 214, 218-219 (C.C.P.A. 1976). Time and difficulty of experiments are not determinative if they are merely routine. M.P.E.P. § 2164.06, page 2100-186.

[2] Applicant respectfully disagrees with the Office's implied assertion that determining probabilities using pre-existing statistical methods is unpredictable. The Office states,"[g]iven the lack of descriptive working examples in the specification, *and the unpredictability of generating probability values,* the specification as filed is not enabling for any method of determining the listed probability values as claimed." Office Action at page 3 (italics added). Applicant respectfully submits that determining probability values using pre-existing statistical methods (*i.e.,* known mathematical equations) is not unpredictable. Applicant respectfully requests that the Office provide legal or other support for the assertion that generating probability values is "unpredictable." Office Action at page 3.

specification need not disclose what is well-known to those skilled in the art and preferably omits that which is well known to those skilled and already available to the public.[3] *See, e.g.,* M.P.E.P. § 2164.05(a), page 2100-185, *citing In re Buchner*, 929 F.2d 660, 661, 18 U.S.P.Q. 2d 1331, 1332 (Fed. Cir. 1991); *Hybritech, Inc. v. Monoclonal Antibodies*, Inc., 802 F.2d 1367, 1384, 231 U.S.P.Q. 81, 94 (Fed. Cir. 1986), cert. denied, 480 U.S. 947 (1987); and *Lindemann Maschinenfabrik GMBH v. American Hoist and Derrick Co.*, 730 F.2d 1452, 1463, 221 U.S.P.Q. 481, 489 (Fed. Cir. 1984).

Applicant respectfully submits that the specification as filed is enabling for the full scope of the claims. The specification describes, and provides working examples for, the use of inhomogeneous Markov models to determine the probabilities for each of the one or more states for a selected nucleotide. *See, e.g.,* specification at pages 19, line 7 though page 27, line 6, and Examples 1 through 3. As such, specification provides sufficient support to enable one of skill in the art, using the state of the art and the specification disclosure, to practice the invention in its full scope without undue experimentation.

Moreover, although Applicant respectfully maintains that no additional information is needed to enable the full scope of the claims, the specification also provides that "[a]ny probability model applicable to nucleic acid sequence state probabilities can be used for the probability steps if the output of the probability model sufficiently supports the method, including inhomogeneous Markov models having fewer than eight states." *See* specification at page 19, lines 22-24. The specification also points that skilled artisan to Durbin *et al.,* Biological Sequence Analysis (1998), described at page 19, lines 26-27 of Applicant's disclosure.[4] Applicant respectfully asserts that for at least these reasons, the specification as filed provides adequate guidance to enable one of

[3] Applicant respectfully submits that the Office has failed to provide any evidence to suggest that the statistical methods taught by Durbin are not well known in the art. Applicant points the Office to the Bibliography of Durbin, which cites over 200 references written by a diversity of authors. Durbin, pages 326-344.

[4] Copies of the Bibliography of Durbin, as well as Durbin Chapters 5 and 11, are enclosed for the Examiner's convenience. *See* Exhibit A.

skill in the art to practice the invention using additional statistical methods that would be substitutable for the four equations that the Office has determined to be enabled.

Moreover, Applicant respectfully disagrees with the Office's implied assertion that the material referred to in Durbin is "essential material." Office Action at page 4; and Office Action mailed January 13, 2003 at page 5. The M.P.E.P. defines "essential material" as including "that which is necessary to provide an enabling disclosure of the claimed invention." M.P.E.P. § 608.01(p), page 600-79.

Applicant respectfully submits that the material in Durbin is not "essential" because it is not necessary to provide an enabling disclosure of the claimed invention. As stated above, disclosure of a single species provides sufficient enabling support if one of skill in the art can, using the state of the art and Applicant's written disclosures, practice the invention in its full scope without undue experimentation. *See In re Wands*, 858 F.2d at 737; *John Hopkins Univ.*, 152 F.3d at 1361; *Spectra-Physics, Inc*, 827 F.2d at 1533; M.P.E.P. § 2164.03. Furthermore, it is well established law that patent applicants are not required to disclose every species enabled by their claims. *See In re Vaeck*, 947 F.2d at 496.

The Office alleges that "Applicant's reliance on prior art methods may only extend to well known methods and that single specific publications do not support their content as being well known." Office Action at pages 4-5. Applicant disagrees. Applicant reiterates that the Office has not provided evidence to suggest that the methods of Durbin are not well known in the art. *See* footnote 3 *infra*. Applicant respectfully submits that the Office has offered no legal support for the assertion that "single specific publications do not support their content as being well known." Furthermore, Applicant's citation of a single reference, rather than a list of references, cannot properly be used as evidence that the information contained therein is not well known in the art. After all, requiring patent applicants to cite a list of cumulative references would contravene the well-known principle that the specification need not disclose what is well-known to those skilled in the art and preferably omits that which is well known to those skilled and already available to the public. M.P.E.P. § 2164.05(a), page 2100-185.

For at least the foregoing reasons, Applicant respectfully asserts that the specification as filed enables a person of skill in the art to practice the invention commensurate in scope with the claims. Applicant respectfully submits that the rejection of claims 1-16 and 41-44 under 35 U.S.C. § 112, first paragraph is improper and should be withdrawn. Reconsideration and withdrawal of these rejections are respectfully requested.

Should the Examiner maintain this rejection based on the contention that the material disclosed in Durbin is not well known to those of ordinary skill in the art, Applicant respectfully requests that the Examiner support this contention by way of affadavit in accordance with 37 C.F.R. § 1.104 (d)(2).

II. Rejections under 35 U.S.C. § 112, Second Paragraph (Indefiniteness)

Claims 1-16 and 41-44 stand rejected under 35 U.S.C. § 112, second paragraph as being allegedly indefinite for failing to particularly point out and distinctly claim the subject matter which Applicant regards as the invention. Office Action at page 5.

(a) Rejection of claims 3 and 11

Claims 3 and 11 stand rejected under 35 U.S.C. § 112, second paragraph on the grounds that they contain mathematical equations that are allegedly confusing as they incorporate "'Φ(f)' representing bias which cancels itself out in each equation, and therefore nullifies its effect on the equation." Office Action at page 5. The Office further alleges that "[i]f the Applicant intends this bias not to represented [sic] by the same exact number in the numerator and denominator, then subscripts, or some other form of notation, would be needed in order to clarify this issue." Office Action at page 5. Applicant respectfully disagrees.

Applicant respectfully disagrees that "Φ(f)" (representing bias) cancels itself out of the equation. Applicant respectfully points out that "Φ(f)" corresponds to a function, and as such, "Φ(f)" can have different numerical values corresponding to different elements in the set of states. *See, e.g.,* specification at page 47, lines 13-20. As acknowledged by the Examiner, when "Φ(f)" has different numerical values

corresponding to different elements in the set of states, "Φ(f)" has different values in the numerator and denominator of the equations in claims 3 and 11, and hence "Φ(f)" does not cancel out. *Compare, e.g.,* calculation at page 46, lines 1-5 with calculation on page 48, lines 5-10. Applicant therefore disagrees that bias cancels itself out of the equation.

Applicant also respectfully disagrees with the Office's assertion that subscripts or other notation are required to clarify this issue. Applicant respectfully submits that acceptability of the claim language depends on whether one of ordinary skill in the art would understand what is claimed, in light of the specification. M.P.E.P. § 2173.05(b). Applicant points out that the specification clearly defines "Φ(f)" as a function. *See, e.g.,* specification at page 24, lines 4-5 and lines 18-25. Applicant respectfully submits that one of skill in the art would understand that a function may be assigned different values under different circumstances, and would also understand that Example 2 illustrates that the values substituted for "Φ(f)" do not cancel out of the equation. *Compare, e.g.,* calculation on page 46, lines 1-5, with calculation on page 48, lines 5-10. Applicant therefore respectfully submits that one of skill in the art would understand the meaning of "Φ(f)" in light of the specification, and that no subscripts or notations are necessary to clarify the issue.

For the foregoing reasons, Applicant respectfully asserts that the specification contains guidelines sufficient to teach the meaning of the claim language "Φ(f)" to one of ordinary skill in the art, and thus, the rejection of claims 3 and 11 under 35 U.S.C. § 112, second paragraph is improper and should be withdrawn. Reconsideration and withdrawal of this rejection is respectfully requested.

(b) Rejection of Claims 1, 7, 8, and 41-44

Claim 1 stands rejected under 35 U.S.C. § 112, second paragraph, on the grounds that it recites the phrase "said probability of said nucleic acid sequence" which is allegedly "vague and indefinite due to the lack of clear antecedent basis for the noted phrase in part d) of claim 1." Office Action at page 5. The Office further alleges that "[t]his lack of antecedent basis and unclear wording is also present in other independent claims 7, 8 (regarding part d) said window probability), 41, 42 (part a) probability of a

window), 43, and 44 (part d) said window probability). This rejection is also applicable to claims 2-6 and 9-16 which are claims dependent from said independent claims due to their direct or indirect dependence." Office Action at page 6. Applicant respectfully disagrees.

Applicant disagrees that there is a lack of clear antecedent basis for the phrase "said probability of said nucleic acid sequence." However, in order to facilitate prosecution, Applicant has amended claim 1.

Applicant further disagrees that there is a lack of antecedent basis and unclear wording in claims 7, 8, and 41-44. Applicant respectfully submits that the specification defines "window" as "a contiguous and defined number of nucleotides within a nucleic acid sequence." *See, e.g.,* Specification at page 17, lines 17-20. Applicant also directs the Office to page 25, line 25-26 of the specification, which states "[i]n order to determine the state probabilities for more than one nucleotide, a window is used for each nucleotide that is examined." Applicant therefore submits that one of ordinary skill, reading the claims in light of the specification and in light of his or her knowledge of the art, would understand the meaning of the phrase "said window probability." When read in light of the specification, the phrase "said window probability" is no less understandable than the phrase "said initial oligonucleotide probability." However, in order to facilitate prosecution, Applicant has amended claims 7, 8, 41, and 44.

Applicant therefore submits that the grounds for the rejection of Claim 1, 7, 8, and 41-44 has been rendered moot. Applicant further submits that the amendments to claims 1, 8, and 41-44 has also rendered moot the rejections of dependent claims 2-6 and 9-16. In light of these remarks, Applicant respectfully requests withdrawal of these rejections.

(c) Rejection of Claims 1, 7, 8, and 41-44

Claims 1, 7, 8, and 41-44 stand rejected under 35 U.S.C. § 112, second paragraph, on the grounds that they recites the phrase "based upon" which allegedly renders unclear "the metes and bounds of the parameters that that determine how much basis is included upon the determinations." Office Action at page 6. The Office further alleges that

"[c]laims 2-6 and 9-16 are also indefinite due to their dependency from claims 1 and 8." Office Action at page 6. Applicant respectfully disagrees.

Applicant disagrees that the phrase "based upon" renders unclear the metes and bounds of the claim. However, in order to facilitate prosecution, Applicant has amended claims 1, 7, 8, and 41-44.

Applicant therefore submits that the grounds for the rejection of Claim 1, 7, 8, and 41-44 has been rendered moot. Applicant further submits that the amendments to claims 1, 8, and 41-44 has also rendered moot the rejections of dependent claims 2-6 and 9-16. In light of these remarks, Applicant respectfully requests withdrawal of these rejections.

(d) Rejection of Claim 7

Claim 7 stands rejected under 35 U.S.C. § 112, second paragraph, on the grounds that it recites the term "capable", which allegedly is a relative term that renders the claim indefinite. Applicant respectfully disagrees.

Even if "capable" were a relative term, the use of a relative term does not make a claim *per se* indefinite. *Seattle Box Co. v. Industrial Crating & Packing, Inc.*, 731 F.2d 818, 826, 221 U.S.P.Q. 568, 574 (Fed. Cir. 1984); M.P.E.P. § 2173.05(b). Breadth in a claim is not to be equated with indefiniteness. *In re Miller*, 441 F.2d 689, 169 U.S.P.Q. 597 (C.C.P..A 1971); M.P.E.P. § 2173.04. The words of a claim must be given their plain meaning unless they are defined in the specification. M.P.E.P. § 2111.01, page 2100-47.

For at least these reasons, Applicant submits that Claim 7 is not indefinite in the recitation of "capable." However, in order to facilitate prosecution, Applicant has amended Claim 7. Applicant therefore submits that the grounds for the rejection of Claim 7 has been rendered moot. In light of these remarks, Applicant respectfully requests withdrawal of this rejection.

(e) Rejection of Claims 3 and 11

Claims 3 and 11 stand rejected under 35 U.S.C. § 112, second paragraph, on the grounds that they are allegedly "vague and indefinite due to the lack of clarity in the

following terms: **f**, **S**, $\mathbf{P_f}$, $\mathbf{P_i}$, and $\mathbf{\Phi}$." Office Action at page 7. The Office further alleges that "a clarification of the metes and bounds is required, by listing in the claim the exact definition of each term in order to make clear whether definitions from the art should be utilized or those in the specification since, as argued by Applicant, art defined (not specification defined) methods are apparently heavily relied upon by Applicant." Office Action at page 7. Applicant respectfully disagrees.

The test for determining whether terms in a given claim are indefinite is whether one skilled in the art would understand what is claimed. *Amgen, Inc. v. Chugai Pharmaceutical Co., Ltd.*, 927 F.2d 1200, 18 U.S.P.Q.2d 1016 (Fed. Cir. 1991), *cert denied,* 112 S.Ct. 169 (1991). M.P.E.P. § 2173.02 states that "[d]efiniteness of claim language must be analyzed, not in a vacuum, but in light of: (A) The content of the particular application disclosure; (B) The teachings of the prior art; and (C) The claim interpretation that would be given by one possessing the ordinary level of skill in the pertinent art at the time the invention was made."

Under M.P.E.P. § 2173.02, the meaning of the terms **f**, **S**, $\mathbf{P_f}$, $\mathbf{P_i}$, and $\mathbf{\Phi}$ must be determined in light of factors (A) through (C) listed above. Applicant respectfully submits that one of skill in the art would understand that **f**, **S**, $\mathbf{P_f}$, $\mathbf{P_i}$, and $\mathbf{\Phi}$ correspond to terms, or parts of terms, of a mathematical equation. Applicant further directs the Office to pages 21-25 of the specification, and Examples 1-2. Applicant respectfully points out that they know of no legal requirement to list "the exact definition of each term" within the claim, and respectfully requests that the Examiner state the legal basis which is relied upon for this statement.

For at least these reasons, Applicant submits that one of ordinary skill in the art, when reading the claim terms **f**, **S**, $\mathbf{P_f}$, $\mathbf{P_i}$, and $\mathbf{\Phi}$ in light of the specification and the teachings of the prior art, would understand what was meant by Claims 3 and 11. Therefore, Applicant respectfully requests that the indefiniteness rejections of claim 3 and 11 under 35 U.S.C. § 112, second paragraph, be withdrawn.

(f) Rejection of Claims 8 and 44

Claims 8 and 44 stand rejected under 35 U.S.C. § 112, second paragraph, on the grounds that they allegedly lack clarity due to the claim language "determining a probability for said window for each of said states. Claims 9-16 are also indefinite due to their dependency from claim 8." Office Action at page 7. Applicant respectfully disagrees.

The Office alleges that "a probability cannot be determined for a window, but rather the states found in the window." Office Action at page 7. Applicant respectfully disagrees. As stated above, under M.P.E.P. § 2173.02, the meaning of the phrase "determining a probability of said window for each of said states" must be determined in light of (A) The content of the particular application disclosure; (B) The teachings of the prior art; and (C) The claim interpretation that would be given by one possessing the ordinary level of skill in the pertinent art at the time the invention was made. Applicant respectfully submits that the specification defines "window" as "a contiguous and defined number of nucleotides within a nucleic acid sequence." *See, e.g.,* Specification at page 17, line 17. Applicant therefore submits that one of ordinary skill, reading this phrase in light of the specification and his or her knowledge of the art, would understand the meaning of the phrase "determining a probability of said window for each of said states." When read in light of the specification, the phrase "determining a probability of said window for each of said states" is no less understandable than the phrase "determining an initial oligonucleotide probability." However, in order to facilitate prosecution, claims 8 and 44 have been amended.

Applicant respectfully submits that, in light of the above arguments, the grounds for the rejection of Claims 8 and 44 has been overcome or rendered moot. Applicant further submits that the rejections of dependent claims 9-16 has also been overcome or rendered moot. In light of these remarks, Applicant respectfully requests withdrawal of these rejections.

III. Rejections under 35 U.S.C. § 102(b)

Claims 1, 4, 5, 7-9, 12, 13, 15, and 41-44 stand rejected under 35 U.S.C. § 102(b) as being allegedly anticipated by Borodovsky *et al.* (Computers Chem., 1993). The Office alleges that "Due to the confusion (see 35 U.S.C. 112, 2nd paragraph rejection above) of "Φ(f)" effectively canceling itself out in the equations of claims 3 and 11, these equations are equivalent to the equations listed on page 129 (Borodovsky *et al.*). Being equivalent equations, if one probability (as provided by Applicant) is "capable of accepting a bias" (claim 7, line 10), then the same probability stated by Borodovsky *et al.* (page 129) must also be capable of accepting a bias. Therefore, Borodovsky *et al.* anticipate the instant invention." Applicant respectfully disagrees.

As noted above, Applicant respectfully disagrees that "Φ(f)" (representing bias) cancels itself out of the equation. Applicant respectfully points out that "Φ(f)" corresponds to a function, and as such, "Φ(f)" can have different numerical values corresponding to different elements in the set of states. *See, e.g.,* specification at page 47, lines 13-20. As acknowledged by the Examiner, when "Φ(f)" has different numerical values corresponding to different elements in the set of states, "Φ(f)" has different values in the numerator and denominator of the equations in claims 3 and 11, and hence "Φ(f)" does not cancel out. For example, Applicant points the Office to Example 2, pages 46-48 of the specification, which illustrates that the values substituted for "Φ(f)" do not cancel out of the equation. *Compare, e.g.,* calculation at page 46, lines 1-5 with calculation on page 48, lines 5-10. Applicant therefore disagrees that bias cancels itself out of the equation.

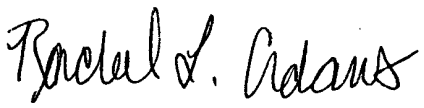
Applicant respectfully disagrees that claims 1, 4, 5, 7-9, 12, 13, 15, and 41-44 are anticipated under §102(b) by Borodovsky. As noted by the Examiner, anticipation under §102(b) requires that every element of a claim appears in a single reference. Applicant respectfully asserts that claims 1, 4, 5, 7-9, 12, 13, 15, and 41-44 each contain the function "Φ(f)" (or the phrase "bias function"), which is lacking in Borodovsky. Applicant respectfully submits that claims 1, 4, 5, 7-9, 12, 13, 15, and 41-44, as amended herein, are therefore not anticipated by Borodovsky. Applicant therefore respectfully requests withdrawal of the rejections under 35 U.S.C. §102(b).

In addition, the Office alleges that "if one claim is 'capable of accepting a bias' (claim 7, line 10), then the same probability stated by Borodovsky *et al.* (page 129) must also be capable of accepting a bias." Office Action at page 8. Applicant respectfully submits that claim 7 has been amended, and that, in light of the amendment of claim 7, the grounds for the rejection of claim 7 have been overcome or rendered moot. In light of these remarks, Applicant respectfully requests withdrawal of the rejection of claim 7.

Conclusion

In view of the above, each of the presently pending claims is believed to be in immediate condition for allowance. Accordingly, the Examiner is respectfully requested to withdraw the outstanding rejections of the claims and to pass this application to issue. The Examiner is encouraged to contact the undersigned at (202) 942-5512 should any additional information be necessary for allowance.

Respectfully submitted,



Rachel L. Adams (Reg. Attorney No. 54,660)
David R. Marsh (Reg. Attorney No. 41,408)
Holly Logue Prutz (Reg. Attorney No. 47,755)

Date: August 12, 2003

ARNOLD & PORTER
555 Twelfth Street, N.W.
Washington, D.C. 20004-1206
(202) 942-5000 telephone
(202) 942-5999 facsimile

Bibliography

Abrahams, J. P., van den Berg, M., van Batenburg, E. and Pleij, C. 1990. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Research* 18:3035–3044.

Allison, L. and Wallace, C. S. 1993. The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimisation of muliple alignments. Technical Report TR 93/188, Monash University Computer Science.

Allison, L., Wallace, C. S. and Yee, C. N. 1992a. Finite-state models in the alignment of macromolecules. *Journal of Molecular Evolution* 35:77–89.

Allison, L., Wallace, C. S. and Yee, C. N. 1992b. Minimum message length encoding, evolutionary trees and multiple alignment. In *Hawaii International Conference on System Sciences*, volume 1, 663–674.

Altschul, S. F. 1989. Gap costs for multiple sequence alignment. *Journal of Theoretical Biology* 138:297–309.

Altschul, S. F. 1991. Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology* 219:555–565.

Altschul, S. F. and Erickson, B. W. 1986. Optimal sequence alignment using affine gap costs. *Bulletin of Mathematical Biology* 48:603–616.

Altschul, S. F. and Gish, W. 1996. Local alignment statistics. *Methods in Enzymology* 266:460–480.

Altschul, S. F. and Lipman, D. J. 1989. Trees, stars, and multiple biological sequence alignment. *SIAM Journal of Applied Mathematics* 49:197–209.

Altschul, S. F., Carroll, R. J. and Lipman, D. J. 1989. Weights for data related by a tree. *Journal of Molecular Biology* 207:647–653.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389–3402.

Asai, K., Hayamizu, S. and Handa, K. 1993. Prediction of protein secondary structure by the hidden Markov model. *Computer Applications in the Biosciences* 9:141–146.

Asmussen, S. 1987. *Applied Probability and Queues*. Wiley.

Atteson, K. 1997. The performance of the neighbor-joining method of phylogeny reconstruction. In Mirkin, B., McMorris, F., Roberts, F. and Rzhetsky, A., eds.,

Mathematical Hierarchies and Biology. American Mathematical Society. 133–148.

Bahl, L. R., Brown, P. F., de Souza, P. V. and Mercer, R. L. 1986. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Proceedings of ICASSP '86*, 49–52.

Bailey, T. L. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In Altman, R., Brutlag, D., Karp, P., Lathrop, R. and Searls, D., eds., *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 28–36. AAAI Press.

Bailey, T. L. and Elkan, C. 1995. The value of prior knowledge in discovering motifs with MEME. In Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. and Wodak, S., eds., *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, 21–29. AAAI Press.

Bairoch, A. and Apweiler, R. 1997. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Research* 25:31–36.

Bairoch, A., Bucher, P. and Hofmann, K. 1997. The PROSITE database, its status in 1997. *Nucleic Acids Research* 25:217–221.

Baldi, P. and Brunak, S. 1998. *Bioinformatics – The Machine Learning Approach*. MIT Press.

Baldi, P. and Chauvin, Y. 1994. Smooth on-line learning algorithms for hidden Markov models. *Neural Computation* 6:307–318.

Baldi, P. and Chauvin, Y. 1995. Protein modeling with hybrid hidden Markov model/neural network architectures. In Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. and Wodak, S., eds., *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, 39–47. AAAI Press.

Baldi, P., Brunak, S., Chauvin, Y. and Krogh, A. 1996. Naturally occurring nucleosome positioning signals in human exons. *Journal of Molecular Biology* 263:503–510.

Baldi, P., Chauvin, Y., Hunkapiller, T. and McClure, M. A. 1994. Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences of the USA* 91:1059–1063.

Bandelt, H.-J. and Dress, A. W. M. 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution* 1:242–252.

Barton, G. J. 1993. An efficient algorithm to locate all locally optimal alignments between two sequences allowing for gaps. *Computer Applications in the Biosciences* 9:729–734.

Barton, G. J. and Sternberg, M. J. E. 1987. A strategy for the rapid multiple alignment of protein sequences. *Journal of Molecular Biology* 198:327–337.

Baserga, S. J. and Steitz, J. A. 1993. The diverse world of small ribonucleoproteins. In Gesteland, R. F. and Atkins, J. F., eds., *The RNA World*. Cold Spring Harbor Press. pp. 359–381.

Bashford, D., Chothia, C. and Lesk, A. M. 1987. Determinants of a protein fold:

unique features of the globin amino acid sequence. *Journal of Molecular Biology* 196:199–216.

Baum, L. E. 1972. An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* 3:1–8.

Bengio, Y., De Mori, R., Flammia, G. and Kompe, R. 1992. Global optimization of a neural network–hidden Markov model hybrid. *IEEE Transactions on Neural Networks* 3:252–259.

Berger, J. O. 1985. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag.

Berger, M. P. and Munson, P. J. 1991. A novel randomized iterative strategy for aligning multiple protein sequences. *Computer Applications in the Biosciences* 7:479–484.

Binder, K. and Heerman, D. W. 1988. *Monte Carlo Simulation in Statistical Mechanics*. Springer-Verlag.

Bird, A. 1987. CpG islands as gene markers in the vertebrate nucleus. *Trends in Genetics* 3:342–347.

Birney, E. and Durbin, R. 1997. Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. In Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C. and Valencia, A., eds., *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, 56–64. AAAI Press.

Bishop, M. J. and Thompson, E. A. 1986. Maximum likelihood alignment of DNA sequences. *Journal of Molecular Biology* 190:159–165.

Borodovsky, M. and McIninch, J. 1993. GENMARK: parallel gene recognition for both DNA strands. *Computers and Chemistry* 17:123–133.

Borodovsky, M. Y., Sprizhitsky, Y. A., Golovanov, E. I. and Alexandrov, A. A. 1986a. Statistical patterns in the primary structure of the functional regions of the Escherichia coli genome. I. Frequency characteristics. *Molecularnaya Biologia* 20:826–833. (English translation).

Borodovsky, M. Y., Sprizhitsky, Y. A., Golovanov, E. I. and Alexandrov, A. A. 1986b. Statistical patterns in the primary structure of the functional regions of the Escherichia Coli genome. II. Nonuniform Markov models. *Molecularnaya Biologia* 20:833–840. (English translation).

Borodovsky, M. Y., Sprizhitsky, Y. A., Golovanov, E. I. and Alexandrov, A. A. 1986c. Statistical patterns in the primary structure of the functional regions of the Escherichia Coli genome. III. Computer recognition of coding regions. *Molecularnaya Biologia* 20:1144–1150. (English translation).

Bowie, J. U., Luthy, R. and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253:164–170.

Box, G. E. P. and Tiao, G. C. 1992. *Bayesian Inference in Statistical Analysis*. Wiley-Interscience.

Branden, C. and Tooze, J. 1991. *Introduction to Protein Structure*. Garland.

Brendel, V., Beckmann, J. S. and Trifonov, E. N. 1986. Linguistics of nucleotide

sequences: morphology and comparison of vocabularies. *Journal of Biomolecular Structure and Dynamics* 4:11–20.

Brooks, D. R. and McLennan, D. A. 1991. *Phylogeny, Ecology and Behaviour.* University of Chicago Press.

Brown, M. and Wilson, C. 1995. RNA pseudoknot modeling using intersections of stochastic context-free grammars with applications to database search. Unpublished manuscript available from http://www.cse.ucsc.edu/research/compbio/pseudoknot.html.

Brown, M., Hughey, R., Krogh, A., Mian, I. S., Sjölander, K. and Haussler, D. 1993. Using Dirichlet mixture priors to derive hidden Markov models for protein families. In Hunter, L., Searls, D. B. and Shavlik, J., eds., *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, 47–55. AAAI Press.

Brunak, S., Engelbrecht, J. and Knudsen, S. 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *Journal of Molecular Biology* 220:49–65.

Bucher, P. and Hofmann, K. 1996. A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system. In States, D. J., Agarwal, P., Gaasterland, T., Hunter, L. and Smith, R. F., eds., *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, 44–51. AAAI Press.

Bucher, P., Karplus, K., Moeri, N. and Hofmann, K. 1996. A flexible motif search technique based on generalized profiles. *Computers and Chemistry* 20:3–24.

Buneman, P. 1971. The recovery of trees from measures of dissimilarity. In Hodson, F. R., Kendall, D. G. and Tautu, P., eds., *Mathematics in the Archaeological and Historical Sciences*. Edinburgh University Press. pp. 387–395.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* 268:78–94.

Camin, J. H. and Sokal, R. R. 1965. A method for deducing branching sequences in phylogeny. *Evolution* 19:311–327.

Cardon, L. R. and Stormo, G. D. 1992. Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *Journal of Molecular Biology* 223:159–170.

Carrillo, H. and Lipman, D. 1988. The multiple sequence alignment problem in biology. *SIAM Journal of Applied Mathematics* 48:1073–1082.

Cary, R. B. and Stormo, G. D. 1995. Graph-theoretic approach to RNA modeling using comparative data. In Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. and Wodak, S., eds., *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, 75–80. AAAI Press.

Casella, G. and Berger, R. L. 1990. *Statistical Inference.* Duxbury Press.

Cavender, J. A. 1978. Taxonomy with confidence. *Mathematical Biosciences* 40:271–280.

Cech, T. R. and Bass, B. L. 1986. Biological catalysis by RNA. *Annual Review of Biochemistry* 55:599–629.

Chan, S. C., Wong, A. K. C. and Chiu, D. K. Y. 1992. A survey of multiple sequence comparison methods. *Bulletin of Mathematical Biology* 54:563–598.

Chang, W. I. and Lawler, E. L. 1990. Approximate string matching in sublinear expected time. In *Proceedings of the 31st Annual IEEE Symposium on Foundations Computer Science*, 116–124. IEEE.

Chao, K. M., Hardison, R. C. and Miller, W. 1994. Recent developments in linear-space alignment methods: a survey. *Journal of Computational Biology* 1:271–291.

Chao, K. M., Pearson, W. R. and Miller, W. 1992. Aligning two sequences within a specified diagonal band. *Computer Applications in the Biosciences* 8:481–487.

Chiu, D. K. Y. and Kolodziejczak, T. 1991. Inferring consensus structure from nucleic acid sequences. *Computer Applications in the Biosciences* 7:347–352.

Chomsky, N. 1956. Three models for the description of language. *IRE Transactions Information Theory* 2:113–124.

Chomsky, N. 1959. On certain formal properties of grammars. *Information and Control* 2:137–167.

Chothia, C. and Lesk, A. M. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO Journal* 5:823–826.

Churchill, G. A. 1989. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* 51:79–94.

Churchill, G. A. 1992. Hidden markov chains and the analysis of genome structure. *Computers and Chemistry* 16:107–115.

Claverie, J.-M. 1994. Some useful statistical properties of position-weight matrices. *Computers and Chemistry* 18:287–294.

Collado-Vides, J. 1989. A transformational-grammar approach to the study of the regulation of gene expression. *Journal of Theoretical Biology* 136:403–425.

Collado-Vides, J. 1991. A syntactic representation of units of genetic information – a syntax of units of genetic information. *Journal of Theoretical Biology* 148:401–429.

Corpet, F. and Michot, B. 1994. RNAlign program: alignment of RNA sequences using both primary and secondary structures. *Computer Applications in the Biosciences* 10:389–399.

Cover, T. M. and Thomas, J. A. 1991. *Elements of Information Theory*. John Wiley & Sons, Inc.

Cox, D. R. 1962. Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, B* 24:406–424.

Cox, D. R. and Miller, H. D. 1965. *The Theory of Stochastic Processes*. Chapman & Hall.

Dandekar, T. and Hentze, M. W. 1995. Finding the hairpin in the haystack: searching for RNA motifs. *Trends in Genetics* 11:45–50.

Dayhoff, M. O., Eck, R. V. and Park, C. M. 1972. In Dayhoff, M. O., ed., *Atlas of Protein Sequence and Structure*, volume 5. National Biomedical Research Foundation, Washington D.C. pp. 89–99.

Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. 1978. A model of evolutionary

change in proteins. In Dayhoff, M. O., ed., *Atlas of Protein Sequence and Structure*, volume 5, supplement 3. National Biomedical Research Foundation, Washington D.C. pp. 345–352.

Dembo, A. and Karlin, S. 1991. Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d. variables. *Annals of Probability* 19:1737–1755.

Dempster, A. P., Laird, N. M. and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* B 39:1–38.

Dong, S. and Searls, D. B. 1994. Gene structure prediction by linguistic methods. *Genomics* 23:540–551.

Doolittle, R. F., Feng, D.-F., Tsang, S., Cho, G. and Little, E. 1996. Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* 271:470–477.

Eck, R. V. and Dayhoff, M. O. 1966. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation.

Eddy, S. R. 1995. Multiple alignment using hidden Markov models. In Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. and Wodak, S., eds., *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, 114–120. AAAI Press.

Eddy, S. R. 1996. Hidden Markov models. *Current Opinion in Structural Biology* 6:361–365.

Eddy, S. R. and Durbin, R. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Research* 22:2079–2088.

Eddy, S. R., Mitchison, G. and Durbin, R. 1995. Maximum discrimination hidden Markov models of sequence consensus. *Journal of Computational Biology* 2:9–23.

Edwards, A. W. F. 1970. Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society, B* 32:155–174.

Edwards, A. W. F. 1992. *Likelihood*. Johns Hopkins Universty Press.

Edwards, A. W. F. 1996. The origin and early development of the method of minimum evolution for the reconstruction of phylogenetic trees. *Systematic Biology* 45:179–191.

Edwards, A. W. F. and Cavalli-Sforza, L. 1963. The reconstruction of evolution. *Annals of Human Genetics* 27:105.

Edwards, A. W. F. and Cavalli-Sforza, L. 1964. Reconstruction of evolutionary trees. In Heywood, V. H. and McNeill, J., eds., *Phenetic and Phylogenetic Classification*. Systematics Association Publication No. 6. pp. 67–76.

Efron, B. and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. Chapman and Hall.

Efron, B., Halloran, E. and Holmes, S. 1996. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences of the USA* 93:13429–13434.

Feller, W. 1971. *An Introduction to Probability Theory and its Applications, Vol II.* John Wiley and Sons.

Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics* 25:471–492.

Felsenstein, J. 1978a. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27:401–410.

Felsenstein, J. 1978b. The number of evolutionary trees. *Systematic Zoology* 27:27–33.

Felsenstein, J. 1981a. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17:368–376.

Felsenstein, J. 1981b. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biological Journal of the Linnean Society* 16:183–196.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791.

Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods in Enzymology* 266:418–427.

Felsenstein, J. and Churchill, G. A. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution* 13:93–104.

Feng, D.-F. and Doolittle, R. F. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution* 25:351–360.

Feng, D.-F. and Doolittle, R. F. 1996. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. *Methods in Enzymology* 266:368–382.

Fichant, G. A. and Burks, C. 1991. Identifying potential tRNA genes in genomic DNA sequences. *Journal of Molecular Biology* 220:659–671.

Fields, D. S. and Gutell, R. R. 1996. An analysis of large rRNA sequences folded by a thermodynamic method. *Folding and Design* 1:419–430.

Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specifed tree topology. *Systematic Zoology* 20:406–416.

Fitch, W. M. and Margoliash, E. 1967a. Construction of phylogenetic trees. *Science* 155:279–284.

Fitch, W. M. and Margoliash, E. 1967b. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. *Biochemical Genetics* 1:65–71.

Frasconi, P. and Bengio, Y. 1994. An EM approach to grammatical inference: input/output HMMs. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, volume 2, 289–294. IEEE Comput. Soc. Press.

Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T. and Turner, D. H. 1986. Improved free-energy parameters for predictions of RNA duplex stability. *Proceedings of the National Academy of Sciences of the USA* 83:9373–9377.

Fujiwara, Y., Asogawa, M. and Konagaya, A. 1994. Stochastic motif extraction using

hidden Markov model. In Altman, R., Brutlag, D., Karp, P., Lathrop, R. and Searls, D., eds., *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 121–129. AAAI Press.

Gautheret, D., Major, F. and Cedergren, R. 1990. Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA. *Computer Applications in the Biosciences* 6:325–331.

Gerstein, M. and Levitt, M. 1996. Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. In States, D. J., Agarwal, P., Gaasterland, T., Hunter, L. and Smith, R. F., eds., *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, 59–67. AAAI Press.

Gerstein, M., Sonnhammer, E. L. L. and Chothia, C. 1994. Volume changes in protein evolution. *Journal of Molecular Biology* 236:1067–1078.

Gersting, J. L. 1993. *Mathematical Structures for Computer Science*. W. H. Freeman.

Gesteland, R. F. and Atkins, J. F., eds. 1993. *The RNA World*. Cold Spring Harbor Laboratory Press.

Gilbert, W. 1986. The RNA world. *Nature* 319:618.

Gold, L., Polisky, B., Uhlenbeck, O. and Yarus, M. 1995. Diversity of oligonucleotide functions. *Annual Review of Biochemistry* 64:763–797.

Goldman, N. 1993. Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* 36:182–198.

Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* 11:725–735.

Gonnet, G. H., Cohen, M. A. and Benner, S. A. 1992. Exhaustive matching of the entire protein sequence database. *Science* 256:1443–1445.

Gotoh, O. 1982. An improved algorithm for matching biological sequences. *Journal of Molecular Biology* 162:705–708.

Gotoh, O. 1993. Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Computer Applications in the Biosciences* 9:361–370.

Gotoh, O. 1996. Significant improvement in accuracy of multiple protein alignments by iterative refinement as assessed by reference to structural alignments. *Journal of Molecular Biology* 264:823–838.

Grate, L. 1995. Automatic RNA secondary structure determination with stochastic context-free grammars. In Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. and Wodak, S., eds., *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, 136–144. AAAI Press.

Gribskov, M. and Veretnik, S. 1996. Identification of sequence patterns with profile analysis. *Methods in Enzymology* 266:198–212.

Gribskov, M., Lüthy, R. and Eisenberg, D. 1990. Profile analysis. *Methods in Enzymology* 183:146–159.

Gribskov, M., McLachlan, A. D. and Eisenberg, D. 1987. Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the USA* 84:4355–4358.

Gultyaev, A. P. 1991. The computer simulation of RNA folding involving pseudoknot formation. *Nucleic Acids Research* 19:2489–2494.

Gumbel, E. J. 1958. *Statistics of Extremes*. Columbia University Press.

Gupta, S. K., Kececioglu, J. D. and Schaffer, A. A. 1995. Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. *Journal of Computational Biology* 2:459–472.

Gutell, R. R. 1993. Collection of small subunit (16S and 16S-like) ribosomal RNA structures. *Nucleic Acids Research* 21:3051–3054.

Gutell, R. R., Power, A., Hertz, G. Z., Putz, E. J. and Stormo, G. D. 1992. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Research* 20:5785–5795.

Hannenhalli, S., Chappey, C., Koonin, E. V. and Pevsner, P. A. 1995. Genome sequence comparison and scenarios for gene rearrangements: a test case. *Genomics* 30:299–311.

Harpaz, Y. and Chothia, C. 1994. Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. *Journal of Molecular Biology* 238:528–539.

Harrison, M. A. 1978. *Introduction to Formal Language Theory*. Addison-Wesley.

Hasegawa, M., Kishino, H. and Yano, T. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22:160–174.

Haussler, D., Krogh, A., Mian, I. S. and Sjölander, K. 1993. Protein modeling using hidden Markov models: analysis of globins. In Mudge, T. N., Milutinovic, V. and Hunter, L., eds., *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*, volume 1, 792–802. IEEE Computer Society Press.

Hebsgaard, S. M., Korning, P. G., Tolstrup, N., Engelbrecht, J., Rouzé, P. and Brunak, S. 1996. Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Research* 24:3439–3452.

Hein, J. 1989a. A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. *Molecular Biology and Evolution* 6:649–668.

Hein, J. 1989b. A tree reconstruction method that is economical in the number of pairwise comparisons used. *Molecular Biology and Evolution* 6:669–684.

Hein, J. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution* 36:396–405.

Henderson, J., Salzberg, S. and Fasman, K. H. 1997. Finding genes in DNA with a hidden Markov model. *Journal of Computational Biology* 4:127–141.

Hendy, M. D. and Penny, D. 1989. A framework for the quantitative study of evolutionary trees. *Systematic Zoology* 38:297–309.

Henikoff, J. G. and Henikoff, S. 1996. Using substitution probabilities to improve

position-specific scoring matrices. *Computer Applications in the Biosciences* 12:135–143.

Henikoff, S. and Henikoff, J. G. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Research* 19:6565–6572.

Henikoff, S. and Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the USA* 89:10915–10919.

Henikoff, S. and Henikoff, J. G. 1994. Position-based sequence weights. *Journal of Molecular Biology* 243:574–578.

Hertz, G. Z., Hartzell III, G. W. and Stormo, G. D. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Computer Applications in the Biosciences* 6:81–92.

Higgins, D. G. and Sharp, P. M. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. *Computer Applications in the Biosciences* 5:151–153.

Higgins, D. G., Bleasby, A. J. and Fuchs, R. 1992. CLUSTAL V: improved software for multiple sequence alignment. *Computer Applications in the Biosciences* 8:189–191.

Hillis, D. M. and Bull, J. J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* 42:182–192.

Hillis, D. M., Bull, J. J., White, M. E., Badgett, M. R. and Molineux, I. J. 1992. Experimental phylogenetics: generation of a known phylogeny. *Science* 255:589–592.

Hirosawa, M., Hoshida, M., Ishikawa, M. and Toya, T. 1993. MASCOT: multiple alignment system for protein sequences based on three-way dynamic programming. *Computer Applications in the Biosciences* 9:161–167.

Hirschberg, D. S. 1975. A linear space algorithm for computing maximal common subsequences. *Communications of the ACM* 18:341–343.

Hogeweg, P. and Hesper, B. 1984. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *Journal of Molecular Evolution* 20:175–186.

Holm, L. and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology* 233:123–138.

Hopcroft, J. E. and Ullman, J. D. 1979. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley.

Huang, X. and Zhang, J. 1996. Methods for comparing a DNA sequence with a protein sequence. *Computer Applications in the Biosciences* 12:497–506.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. In Futuyma, D. and Antonovics, J., eds., *Gene Genealogies and the Coalescent Process*. Oxford University Press. pp. 1–44.

Huelsenbeck, J. P. and Rannala, B. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* 276:227–232.

Hughey, R. and Krogh, A. 1996. Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Computer Applications in the Biosciences* 12:95–107.

Jacob, F. 1977. Evolution and tinkering. *Science* 196:1161–1166.

Jefferys, W. H. and Berger, J. O. 1992. Ockham's razor and Bayesian analysis. *American Scientist* 80:64–72.

Juang, B. H. and Rabiner, L. R. 1991. Hidden Markov models for speech recognition. *Technometrics* 33:251–272.

Jukes, T. H. and Cantor, C. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism*. Academic Press. pp. 21–132.

Karlin, S. and Altschul, S. F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the USA* 87:2264–2268.

Karlin, S. and Altschul, S. F. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Sciences of the USA* 90:5873–5877.

Karplus, K. 1995. Evaluating regularizers for estimating distributions of amino acids. In Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. and Wodak, S., eds., *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, 188–196. AAAI Press.

Keeping, E. S. 1995. *Introduction to Statistical Inference*. Dover Publications.

Kim, J. and Pramanik, S. 1994. An efficient method for multiple sequence alignment. In Altman, R., Brutlag, D., Karp, P., Lathrop, R. and Searls, D., eds., *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 212–218. AAAI Press.

Kim, J., Pramanik, S. and Chung, M. J. 1994. Multiple sequence alignment using simulated annealing. *Computer Applications in the Biosciences* 10:419–426.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of necleotide sequences. *Journal of Molecular Evolution* 16:111–120.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kingman, J. F. C. 1982a. The coalescent. *Stochastic Processes and their Applications* 13:235–248.

Kingman, J. F. C. 1982b. On the genealogy of large populations. *Journal of Applied Probability* 19A:27–43.

Kirkpatrick, S., Gelatt, Jr., C. D. and Vecchi, M. P. 1983. Optimization by simulated annealing. *Science* 220:671–680.

Kishino, H., Miyata, T. and Hasegawa, M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution* 31:151–160.

Konings, D. A. M. and Gutell, R. R. 1995. A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *RNA* 1:559–574.

Konings, D. A. M. and Hogeweg, P. 1989. Pattern analysis of RNA secondary structure: similarity and consensus of minimal-energy folding. *Journal of Molecular Biology* 207:597–614.

Krogh, A. 1994. Hidden Markov models for labeled sequences. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, 140–144. IEEE Computer Society Press.

Krogh, A. 1997a. Gene finding: putting the parts together. In Bishop, M., ed., *Guide to Human Genome Computing*. Academic Press, 2nd edition. To appear.

Krogh, A. 1997b. Two methods for improving performance of a HMM and their application for gene finding. In Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C. and Valencia, A., eds., *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, 179–186. AAAI Press.

Krogh, A. 1998. An introduction to hidden Markov models for biological sequences. In Salzberg, S., Searls, D. and Kasif, S., eds., *Computational Biology: Pattern Analysis and Machine Learning Methods*. Elsevier. Chapter 4. In press.

Krogh, A. and Mitchison, G. 1995. Maximum entropy weighting of aligned sequences of proteins or DNA. In Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. and Wodak, S., eds., *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, 215–221. AAAI Press.

Krogh, A., Mian, I. S. and Haussler, D. 1994. A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Research* 22:4768–4778.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K. and Haussler, D. 1994. Hidden Markov models in computational biology: applications to protein modeling. *Journal of Molecular Biology* 235:1501–1531.

Kuhner, M. K., Yamato, J. and Felsenstein, J. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics* 140:1421–1430.

Kulp, D., Haussler, D., Reese, M. G. and Eeckman, F. H. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. In States, D. J., Agarwal, P., Gaasterland, T., Hunter, L. and Smith, R. F., eds., *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, 134–142. AAAI Press.

Langley, C. H. and Fitch, W. M. 1974. An examination of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution* 3:161–177.

Lari, K. and Young, S. J. 1990. The estimation of stochastic context-free grammars using the inside–outside algorithm. *Computer Speech and Language* 4:35–56.

Lari, K. and Young, S. J. 1991. Applications of stochastic context-free grammars using the inside–outside algorithm. *Computer Speech and Language* 5:237–257.

Larsen, N. and Zwieb, C. 1993. The signal recognition particle database (SRPDB). *Nucleic Acids Research* 21:3019–3020.

Law, A. M. and Kelton, W. D. 1991. *Simulation Modelling and Analysis*. McGraw-Hill.

Lawrence, C. E. and Reilly, A. A. 1990. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* 7:41–51.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. and Wootton, J. C. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262:208–214.

Lefebvre, F. 1995. An optimized parsing algorithm well suited to RNA folding. In Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. and Wodak, S., eds., *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, 222–230. AAAI Press.

Lefebvre, F. 1996. A grammar-based unification of several alignment and folding algorithms. In States, D. J., Agarwal, P., Gaasterland, T., Hunter, L. and Smith, R. F., eds., *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, 143–154. AAAI Press.

Lindenmayer, A. 1968. Mathematical models for cellular interactions in development I. filaments with one-sided inputs. *Journal of Theoretical Biology* 18:280–299.

Lipman, D. J., Altschul, S. F. and Kececioglu, J. D. 1989. A tool for multiple sequence alignment. *Proceedings of the National Academy of Sciences of the USA* 86:4412–4415.

Lisacek, F., Diaz, Y. and Michel, F. 1994. Automatic identification of group I intron cores in genomic DNA sequences. *Journal of Molecular Biology* 235:1206–1217.

Lowe, T. M. and Eddy, S. R. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* 25:955–964.

Lukashin, A. V., Engelbrecht, J. and Brunak, S. 1992. Multiple alignment using simulated annealing: branch point definition in human mRNA splicing. *Nucleic Acids Research* 20:2511–2516.

Luthy, R., McLachlan, A. D. and Eisenberg, D. 1991. Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins* 10:229–239.

Luthy, R., Xenarios, I. and Bucher, P. 1994. Improving the sensitivity of the sequence profile method. *Protein Science* 3:139–146.

MacKay, D. J. C. 1992. Bayesian interpolation. *Neural Computation* 4:415–447.

MacKay, D. J. C. and Peto, L. 1995. A hierarchical Dirichlet language model. *Natural Language Engineering* 1:1–19.

Margalit, H., Shapiro, B. A., Oppenheim, A. B. and Maizel, J. V. 1989. Detection of common motifs in RNA secondary structures. *Nucleic Acids Research* 17:4829–4845.

Mathews, J. and Walker, R. L. 1970. *Mathematical Methods of Physics*. W. A. Benjamin.

Mau, B., Newton, M. A. and Larget, B. 1996. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Technical Report 961, Statistics Department, University of Wisconsin-Madison.

Maxwell, E. S. and Fournier, M. J. 1995. The small nucleolar RNAs. *Annual Review of Biochemistry* 64:897–934.

McCaskill, J. S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29:1105–1119.

McClure, M. A., Vasi, T. K. and Fitch, W. M. 1994. Comparative analysis of multiple protein-sequence alignment methods. *Journal of Molecular Evolution* 11:571–592.

McKeown, M. 1992. Alternative mRNA splicing. *Annual Review of Cell Biology* 8:133–155.

Melefors, O. and Hentze, M. W. 1993. Translational regulation by mRNA/protein interactions in eukaryotic cells: ferritin and beyond. *BioEssays* 15:85–90.

Meng, X.-L. and Rubin, D. B. 1992. Recent extensions to the EM algorithm. *Bayesian Statistics* 4:307–320.

Mevissen, H. T. and Vingron, M. 1996. Quantifying the local reliability of a sequence alignment. *Protein Engineering* 9:127–132.

Miller, W. and Myers, E. W. 1988. Sequence comparison with concave weighting functions. *Bulletin of Mathematical Biology* 50:97–120.

Mitchison, G. 1998. Probabilistic modelling of phylogeny and alignment. *Molecular Biology and Evolution* submitted.

Mitchison, G. and Durbin, R. 1995. Tree-based maximal likelihood substitution matrices and hidden Markov models. *Journal of Molecular Evolution* 41:1139–1151.

Miyazawa, S. 1994. A reliable sequence alignment method based on probabilities of residue correspondence. *Protein Engineering* 8:999–1009.

Mott, R. 1992. Maximum likelihood estimation of the statistical distribution of Smith–Waterman local sequence similarity scores. *Bulletin of Mathematical Biology* 54:59–75.

Myers, E. W. 1994. A sublinear algorithm for approximate keyword searching. *Algorithmica* 12:345–374.

Myers, E. W. and Miller, W. 1988. Optimal alignments in linear space. *Computer Applications in the Biosciences* 4:11–17.

Myers, G. 1995. Approximately matching context-free languages. *Information Processing Letters* 54:85–92.

Neal, R. M. 1996. *Bayesian Learning in Neural Networks.* Springer (Lecture Notes in Statistics).

Neal, R. M. and Hinton, G. E. 1993. A new view of the EM algorithm that justifies incremental and other variants. Preprint, Dept. of Computer Science, Univ. of Toronto, available from ftp://archive.cis.ohio-state.edu/pub/neuroprose/neal.em.ps.Z.

Needleman, S. B. and Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48:443–453.

Noller, H. F., Hoffarth, V. and Zimniak, L. 1992. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science* 256:1416–1419.

Normandin, Y. and Morgera, S. D. 1991. An improved MMIE training algorithm for speaker-independent, small vocabulary, continuous speech recognition. In *Proceedings of ICASSP '91*, 537–540.

Nussinov, R., Pieczenik, G., Griggs, J. R. and Kleitman, D. J. 1978. Algorithms for loop matchings. *SIAM Journal of Applied Mathematics* 35:68–82.

Pavesi, A., Conterlo, F., Bolchi, A., Dieci, G. and Ottonello, S. 1994. Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight

matrix analysis of transcriptional control regions. *Nucleic Acids Research* 22:1247–1256.

Pearson, W. R. 1995. Comparison of methods for searching protein sequence databases. *Protein Science* 4:1145–1160.

Pearson, W. R. 1996. Effective protein sequence comparison. *Methods in Enzymology* 266:227–258.

Pearson, W. R. and Lipman, D. J. 1988. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the USA* 4:2444–2448.

Pearson, W. R. and Miller, W. 1992. Dynamic programming algorithms for biological sequence comparison. *Methods in Enzymology* 210:575–601.

Pedersen, A. G., Baldi, P., Brunak, S. and Chauvin, Y. 1996. Characterization of prokaryotic and eukaryotic promoters using hidden Markov models. In States, D. J., Agarwal, P., Gaasterland, T., Hunter, L. and Smith, R. F., eds., *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, 182–191. AAAI Press.

Peltz, S. W. and Jacobson, A. 1992. mRNA stability: in trans-it. *Current Opinion in Cell Biology* 4:979–983.

Pesole, G., Attimonelli, M. and Saccone, C. 1994. Linguistic approaches to the analysis of sequence information. *Trends in Biotechnology* 12:401–408.

Pietrokovski, S., Hirshon, J. and Trifonov, E. N. 1990. Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. *Journal of Biomolecular Structure and Dynamics* 7:1251–1268.

Preparata, F. P. and Shamos, M. I. 1985. *Computational Geometry*. Springer-Verlag.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. 1992. *Numerical Recipes in C*. Cambridge University Press.

Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77:257–286.

Rabiner, L. R. and Juang, B. H. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine* 3:4–16.

Rabiner, L. R. and Juang, B. H. 1993. *Fundamentals of Speech Recognition*. Prentice-Hall.

Rannala, B. and Yang, Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution* 43:304–311.

Reese, M. G., Eeckman, F. H., Kulp, D. and Haussler, D. 1997. Improved splice site detection in Genie. *Journal of Computational Biology* 4:311–323.

Renals, S., Morgan, N., Bourlard, H., Cohen, M. and Franco, H. 1994. Connectionist probability estimators in hmm speech recognition. *IEEE Transactions on Speech and Audio Processing* 2:161–174.

Riis, S. K. and Krogh, A. 1997. Hidden neural networks: a framework for HMM/NN hybrids. In *Proceedings of ICASSP '97*, 3233–3236. IEEE.

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press.

Rosenblueth, D. A., Thieffry, D., Huerta, A. M., Salgado, H. and Collado-Vides, J. 1996. Syntactic recognition of regulatory regions in *Escherichia coli*. *Computer Applications in the Biosciences* 12:415–422.

Russell, R. B. and Barton, G. J. 1992. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* 14:309–323.

Saitou, N. 1996. Reconstruction of gene trees from sequence data. *Methods in Enzymology* 266:427–448.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406–425.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjölander, K., Underwood, R. C. and Haussler, D. 1994. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research* 22:5112–5120.

Sankoff, D. 1975. Minimal mutation trees of sequences. *SIAM Journal of Applied Mathematics* 28:35–42.

Sankoff, D. and Cedergren, R. J. 1983. Simultaneous comparison of three or more sequences related by a tree. In Sankoff, D. and Kruskal, J. B., eds., *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*. Addison-Wesley. Chapter 9, pp. 253–264.

Sankoff, D. and Kruskal, J. B. 1983. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley.

Sankoff, D., Morel, C. and Cedergren, R. J. 1973. Evolution of 5S RNA and the nonrandomness of base replacement. *Nature New Biology* 245:232–234.

Schneider, T. D. and Stephens, R. M. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* 18:6097–6100.

Schuster, P. 1995. How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *Journal of Biotechnology* 41:239–257.

Schuster, P., Fontana, W., Stadler, P. F. and Hofacker, I. L. 1994. From sequences to shapes and back: a case study in RNA secondary structures. *Proceedings of the Royal Society: Biological Sciences, Series B* 255:279–284.

Schwartz, R. and Chow, Y.-L. 1990. The N-best algorithm: an efficient and exact procedure for finding the n most likely hypotheses. In *Proceedings of ICASSP'90*, 81–84.

Searls, D. B. 1992. The linguistics of DNA. *American Scientist* 80:579–591.

Searls, D. B. and Murphy, K. P. 1995. Automata-theoretic models of mutation and alignment. In Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. and Wodak, S., eds., *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, 341–349. AAAI Press.

Shapiro, B. A. and Wu, J. C. 1996. An annealing mutation operator in the genetic algorithms for RNA folding. *Computer Applications in the Biosciences* 12:171–180.

Shapiro, B. A. and Zhang, K. 1990. Comparing multiple RNA secondary structures using tree comparisons. *Computer Applications in the Biosciences* 6:309–318.

Shimamura, M., Yasue, H., Ohshima, K., Abe, H., Kato, H., Kishiro, T., Goto, M.,

Munechika, I. and Okada, N. 1997. Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature* 388:666–670.

Shpaer, E. G., Robinson, M., Yee, D., Candlin, J. D., Mines, R. and Hunkapiller, T. 1996. Sensitivity and selectivity in protein similarity searches: a comparison of Smith–Waterman in hardware to BLAST and FASTA. *Genomics* 38:179–191.

Sibbald, P. R. and Argos, P. 1990. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *Journal of Molecular Biology* 216:813–818.

Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S. and Haussler, D. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Computer Applications in the Biosciences* 12:327–345.

Smith, T. F. and Waterman, M. S. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* 147:195–197.

Sokal, R. R. and Michener, C. D. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* 28:1409–1438.

Sonnhammer, E. L. L., Eddy, S. R. and Durbin, R. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28:405–420.

Staden, R. 1988. Methods to define and locate patterns of motifs in sequences. *Computer Applications in the Biosciences* 4:53–60.

Steinberg, S., Misch, A. and Sprinzl, M. 1993. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Research* 21:3011–3015.

Stolcke, A. and Omohundro, S. M. 1993. Hidden Markov model induction by Bayesian model merging. In Hanson, S. J., Cowan, J. D. and Giles, C. L., eds., *Advances in Neural Information Processing Systems 5*, volume 5, 11–18. Morgan Kaufmann Publishers, Inc.

Stormo, G. D. 1990. Consensus patterns in DNA. *Methods in Enzymology* 183:211–221.

Stormo, G. D. and Hartzell III, G. W. 1989. Identifying protein–binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Sciences of the USA* 86:1183–1187.

Stormo, G. D. and Haussler, D. 1996. Optimally parsing a sequence into different classes based on multiple types of evidence. In States, D. J., Agarwal, P., Gaasterland, T., Hunter, L. and Smith, R. F., eds., *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, 369–375. AAAI Press.

Studier, J. A. and Keppler, K. J. 1988. A note on the neighbour-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution* 5:729–731.

Swofford, D. L. and Olsen, G. J. 1996. Phylogeny reconstruction. In Hillis, D. M. and Moritz, C., eds., *Molecular Systematics*. Sinauer Associates. pp. 407–511.

Tatusov, R. L., Altschul, S. F. and Koonin, E. V. 1994. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proceedings of the National Academy of Sciences of the USA* 91:12091–12095.

Taylor, W. R. 1987. Multiple sequence alignment by a pairwise algorithm. *Computer Applications in the Biosciences* 3:81–87.

Thompson, E. A. 1975. *Human Evolutionary Trees*. Cambridge University Press.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. 1994a. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22:4673–4680.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. 1994b. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Computer Applications in the Biosciences* 10:19–29.

Thorne, J. L., Kishino, H. and Felsenstein, J. 1992. Inching toward reality: an improved likelihood model of sequence evolution. *Methods in Enzymology* 34:3–16.

Tolstrup, N., Rouzé, P. and Brunak, S. 1997. A branch point consensus from Arabidopsis found by non-circular analysis allows for better prediction of acceptor sites. *Nucleic Acids Research* 25:3159–3164.

Tuerk, C., MacDougal, S. and Gold, L. 1992. RNA pesudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proceedings of the National Academy of Sciences of the USA* 89:6988–6992.

Turner, D. H., Sugimoto, N., Jaeger, J. A., Longfellow, C. E., Freier, S. M. and Kierzek, R. 1987. Improved parameters for prediction of RNA structure. *Cold Spring Harbor Symposia Quantitative Biology* 52:123–133.

van Batenburg, F. H. D., Gultyaev, A. P. and Pleij, C. W. A. 1995. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *Journal of Theoretical Biology* 174:269–280.

Vingron, M. 1996. Near-optimal sequence alignment. *Current Opinion in Structural Biology* 6:346–352.

Vingron, M. and Waterman, M. S. 1994. Sequence alignment and penalty choice: review of concepts, case studies and implications. *Journal of Molecular Biology* 235:1–12.

Waterman, M. S. 1995. *Introduction to Computational Biology*. Chapman & Hall.

Waterman, M. S. and Eggert, M. 1987. A new algorithm for best subsequence alignments with application to tRNA–rRNA comparisons. *Journal of Molecular Biology* 197:723–725.

Waterman, M. S. and Perlwitz, M. D. 1984. Line geometries for sequence comparisons. *Bulletin of Mathematical Biology* 46:567–577.

Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A. and Weiner, A. M. 1987. *Molecular Biology of the Gene*. Benjamin/Cummings.

Wilmanns, M. and Eisenberg, D. 1993. Three-dimensional profiles from residue-pair preferences: identification of sequences with beta/alpha-barrel fold. *Proceedings of the National Academy of Sciences of the USA* 90:1379–1383.

Witherell, G. W., Gott, J. M. and Uhlenbeck, O. C. 1991. Specific interaction between RNA phage coat proteins and RNA. *Progress in Nucleic Acid Research and Molecular Biology* 40:185–220.

Woese, C. R. and Pace, N. R. 1993. Probing RNA structure, function, and history by comparative analysis. In Gesteland, R. F. and Atkins, J. F., eds., *The RNA World*. Cold Spring Harbor Laboratory Press. pp. 91–117.

Wray, G. A., Levinto, J. S. and Shapiro, L. H. 1996. Molecular evidence for deep precambrian divergences among metazoan phyla. *Science* 274:568–573.

Wu, S. and Manber, U. 1992. Fast text searching allowing errors. *Communications of the ACM* 35:83–90.

Yada, T. and Hirosawa, M. 1996. Detection of short protein coding regions within the Cyanobacterium genome: application of the hidden Markov model. *DNA Research* 3:355–361.

Yada, T., Sazuka, T. and Hirosawa, M. 1997. Analysis of sequence patterns surrounding the translation initiation sites on Cyanobacterium genome using the hidden Markov model. *DNA Research* 4:1–7.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* 10:1396–1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* 39:306–314.

Zuckerkandel, E. and Pauling, L. 1962. Molecular disease, evolution and genetic heterogeneity. In Marsha, M. and Pullman, B., eds., *Horizons in Biochemistry*. Academic Press. pp. 189–225.

Zuker, M. 1989a. Computer prediction of RNA structure. *Methods in Enzymology* 180:262–288.

Zuker, M. 1989b. On finding all suboptimal foldings of an RNA molecule. *Science* 244:48–52.

Zuker, M. 1991. Suboptimal sequence alignment in molecular biology: alignment with error analysis. *Journal of Molecular Biology* 221:403–420.

Zuker, M. and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research* 9:133–148.